Machine Learning 6.867 - Project

December 10, 2015

1 Introduction

Logistic regression is a widely-used method for classification. Given training data $(\mathbf{x}_i, y_i)_{i=1}^n$ with features $\mathbf{x}_i \in \mathbb{R}^p$ and labels $y_i \in \{-1, +1\}$, we would like to find $P(y = 1|\mathbf{x})$ for an arbitrary vector $\mathbf{x}$. For logistic regression, we assume a logistic model, where $P(y = 1|\mathbf{x}) = \dfrac{1}{1 + e^{-y\mathbf{w}^T\mathbf{x}}}$, and $\mathbf{w} \in \mathbb{R}^p$ is a weight vector. To determine the value of the weight vector, we can do maximum-likelihood estimation, and optimize to find the value of $\mathbf{w}$ which maximizes the negative log-likelihood. This results in a concave, nonlinear optimization problem solvable via gradient descent.

Alternatively, we might take the Bayesian perspective and assume a prior on $\mathbf{w}$. This is preferable for cases where we interested in the covariance of the weight vector and/or the predicted probabilities of new data points. However, this is a difficult problem and there is no simple analytic formula to compute the posterior and predictive distributions directly. Therefore, in this case we turn to variational methods in order to obtain a close approximation to the posterior for $\mathbf{w}$. From this, we can then obtain estimates for the predictive distribution $P(y = 1|\mathbf{x})$.

In this paper, we develop and test an implementation of Mean-Field Variational Bayes logistic regression based on previous literature. Paper [1] provides a walkthrough of MFVB applied to logistic regression, and sample MATLAB code is available on Github. We develop our own implementation based off of this package of code, and we code these functions in Julia. Going beyond this work, we compare our method against Markov-Chain Monte-Carlo, and implemented this method using an existing package in R. We performed sensitivity analysis varying the hyperpriors of $\mathbf{w}$ to generate the data, and we obtain a variety of different simulated datasets. In Section 5, we present train and test set accuracy for all of the simulated datasets, for standard logistic regression, MFVB, and MCMC. We obtain estimates for the predicted probabilies of MCMC by taking the average of the weight vector $\mathbf{w}$ over all iterations, and then use that to compute the Bernoulli probabilities. In addition, we also present detailed MCMC results and comparisons to contour plots of the MFVB logistic regression weight posteriors for Dataset 0 in Section 4.

In Section 2, we introduce Mean-Field Variational Bayes applied to the problem of logistic regression. In Section 3, we describe our code implementation of MFVB logistic regression and the simulated datasets we generated to test our methods. In Section 4, we describe the Markov-Chain Monte Carlo method that we used as a benchmark to evaluate the posterior distribution. In Section 5, we present the out-of-sample accuracy results for all of the methods on the simulated datasets, and in Section 6, we describe our goals for future work.

2 Mean-Field Variational Bayes

Mean-field variational Bayes (MFVB) is a method for approximating the posterior distribution. In general, we have unknown parameters $w_1, w_2, \ldots, w_n$ that we have priors on, and our objective is to find the joint distribution $p(w_1, w_2, \ldots, w_n)$. Assuming that our approximate distribution is in the family $Q = \{q : q(w_1, w_2, \ldots, w_n) = q(w_1)q(w_2)\ldots q(w_n)\}$, we find $q^* \in Q$ that minimizes the KL-divergence with p, i.e. $q^* = \min KL(q||p)$. In particular, for logistic regression, the analytical form of the posterior is unknown and has been approximated with MFVB in the literature [2]. We use local variational bounds on the conditional probability using the convexity of the logarithm function. In particular, we use a variational treatment based on the approach of Jaakkola and Jordan (2000). This approach consists of approximation the likelihood function of the logistic regression, governed by the sigmoid function, by the exponential of the a quadratic form, leading to a gaussian approximation of the posterior distribution. More explicitly, if $y \in \{-1, 1\}$ is a target variable for a data vector x then the likelihood function of the target variable y is:

$$p(y|x, w) = \sigma(yw^T x) \tag{1}$$

with w being the logistic regression weight, and $\sigma(x) = \dfrac{1}{1 + \exp(-x)}$ the sigmoid function. Using a transformation of a the logarithm of the sigmoid and the concept of convex duality, we get:

$$\sigma(x) \geq \sigma(\xi)\exp((x - \xi)/2 - \lambda(\xi)(x^2 - \xi^2)) \tag{2}$$

where

$$\lambda(\xi) = \frac{1}{2\xi}[\sigma(\xi) - \frac{1}{2}] = \frac{1}{4\xi}\tanh(\frac{\xi}{2})$$

and ξ is a variational parameter.

Therefore, if we let $a = w^T x$ we get:

$$p(y|x, w) \geq e^{ya}\sigma(\xi)\exp\{-(\xi + a)/2 - \lambda(\xi)(a^2 - \xi^2)\} \tag{3}$$

To every training set observation (x_n, y_n), there is a variational parameter ξ_n associated. We apply the bound above to each of the terms in the likelihood function. Let $Y = [y_1, y_2, \ldots, y_n]^T$ and X be the data matrix. Then the likelihood function is:

$$p(Y|X, w) = \prod_{i=1}^{N} p(y_i|x_i, w) = \prod_{i=1}^{N} \sigma(yw^T x) \tag{4}$$

and thus we obtain the following bound on the marginal data likelihood:

$$p(Y|X, w) \geq h(w, \xi) \tag{5}$$

and

$$h(w, \xi) = \prod_{i=1}^{N} e^{y_i w^T x_i}\sigma(\xi_i)\exp\{-(\xi_i + w^T x_i)/2 - \lambda(\xi_i)((w^T x_i)^2 - {\xi_i}^2)\}$$

This approximation is used because the sigmoid data likelihood does not have a conjugate in the exponential family of priors. The approximation we yielded is quadratic in w in the exponential, and we use the conjugate gaussian prior:

$$p(w|\alpha) = \mathcal{N}(0, \alpha^{-1}\mathcal{I}) \tag{6}$$

and we can also model the hyper-parameter α with a conjugate Gamma distribution:

$$p(\alpha) = Gam(\alpha|a_0, b_0) \tag{7}$$

Variational Bayesian inference aims at maximizing a lower bound of the data log-likelihood. The log-likelihood is

$$\ln p(Y|X) = \ln \int\int p(Y|X, w)p(w|\alpha)p(\alpha)dwd\alpha \tag{8}$$

We approximate the posterior $p(w, \alpha|X)$ by the variational distribution $Q(w, \alpha)$ that can be factorized to $Q(\alpha)Q(w)$. The bound of the log-likelihood is as follows:

$$\ln p(Y|X) \geq \mathcal{L}(Q) = \ln \int\int Q(w, \alpha)\ln\frac{p(Y|X, w)p(w|\alpha)p(\alpha)}{Q(w, \alpha)}dwd\alpha \tag{9}$$

Hence, using Equation (5), obtain a variational bound $\tilde{\mathcal{L}}(Q, \xi)$ that we aim at maximizing:

$$\tilde{\mathcal{L}}(Q, \xi) = \int\int Q(w, \alpha) \ln \frac{h(w, \xi)p(w|\alpha)p(\alpha)}{Q(w, \alpha)} dw d\alpha \tag{10}$$

After we substitute $Q(w, \alpha)$ with $Q(\alpha)Q(w)$ and we calculate the expectations of *alpha* and w as in the general MFVB, we obtain this expression of $\tilde{\mathcal{L}}(Q, \xi)$:

$$\tilde{\mathcal{L}}(Q, \xi) = \frac{1}{2} w_N^T V_N^{-1} w_N + \frac{1}{2} \ln |V_N| + \sum_n \left(\ln \sigma(\xi_n) - \frac{\xi_n}{2} + \lambda(\xi_n)\xi_n^2 \right) - \ln \Gamma(a_0) + a_0 \ln(b_0) - b_0 \frac{a_N}{b_N} - a_N \ln(b_N) - \ln \Gamma(a_N) + a_N \tag{11}$$

with

$$a_N = a_0 + \frac{D}{2}$$

$$b_N = b_0 + \frac{1}{2} w_N^T w_N + Tr(V_N)$$

$$V_N^{-1} = \frac{a_N}{b_N} \mathcal{I} + 2 \sum_n \lambda(\xi_n) x_n x_n^T$$

$$w_N = V_N \sum_n \frac{y_n}{2} x_n$$

and

$$Q^*(w) = \mathcal{N}(w|w_N, V_N)$$

$$Q^*(\alpha) = Gam(\alpha|a_N, b_N)$$

This bound depends on ξ. We maximize this bound with respect to ξ and we find:

$$(\xi_n^{new})^2 = x_n^T (V_N + w_N w_N^T) x_N \tag{12}$$

We use the EM algorithm to update the equations for $w_N, V_N, a_N, b_N,$ and ξ in order to maximize the variational bound $\tilde{\mathcal{L}}(Q, \xi)$, until it reaches a plateau.

3 Implementation

3.1 MFVB Logistic Regression

We wrote a batch implementation of function that computes the variational posterior parameters w, V, V^{-1} by updating all of the ξ_n's at once. Its input were the hyper-priors a_0, b_0. The values we took are 10^{-2} and 10^{-4} respectively.

3.2 Simulated Datasets

To have different datasets, we used a fixed data vector matrix $\mathbf{X}$ and we took 10 random values of a_0 and b_0 each, from which we sampled the label vectors $\mathbf{y}$ using the prior and hyperprior distributions listed in the MFVB section. We thus had 10 different datasets, in the sense that every dataset had different labels for the points $\mathbf{X}$.

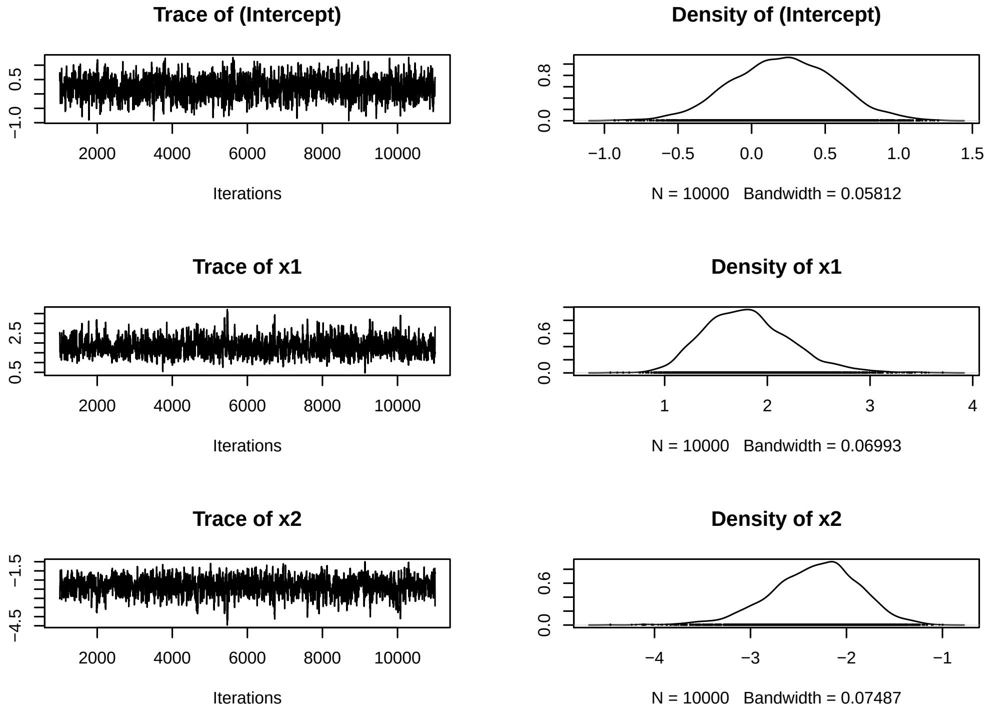

Figure 1: MCMC simulations of logistic regression weights for Dataset 0, and corresponding marginal density plots, assuming an improper uniform prior. 10,000 iterations total.

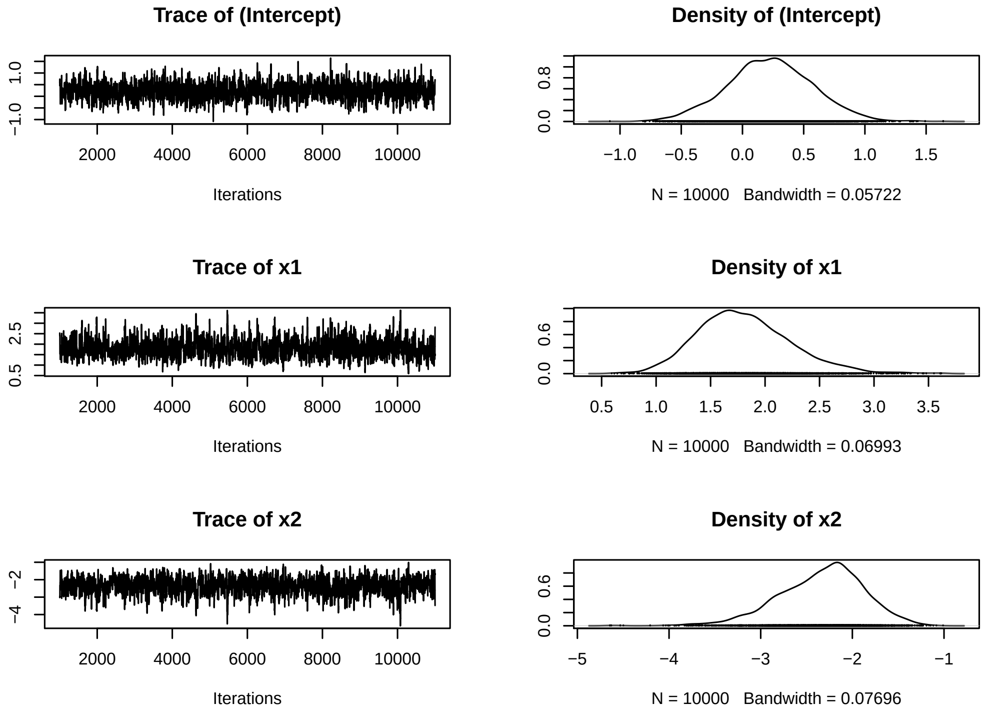

Figure 2: MCMC simulations of logistic regression weights for Dataset 0, and corresponding marginal density plots, assuming a normal prior $\mathcal{N}(\mathbf{0}, 1000I)$. 10,000 iterations total.

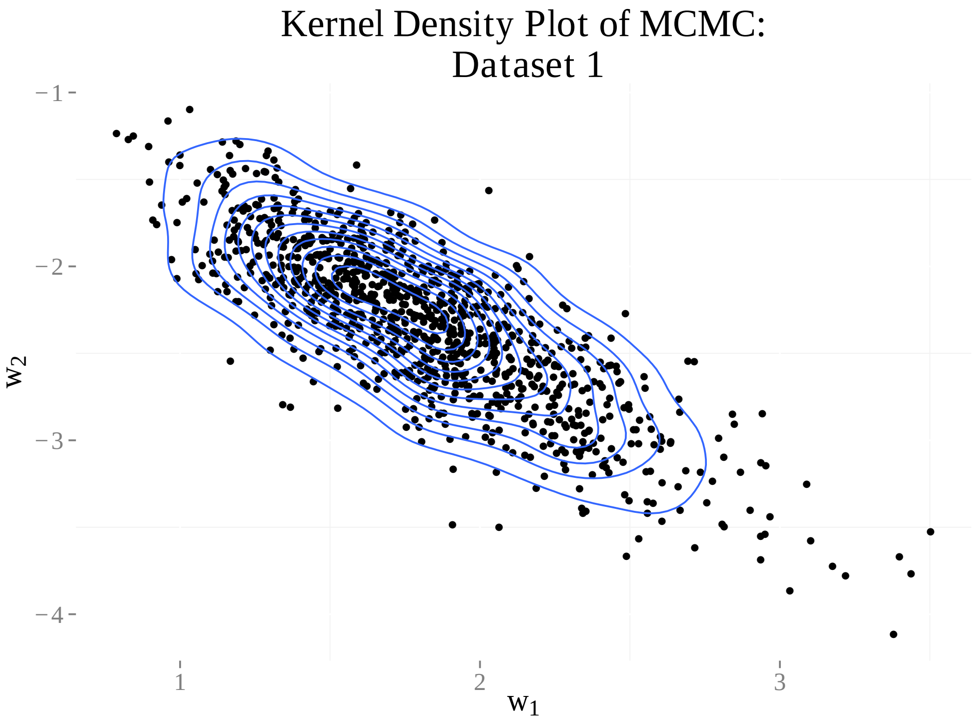

Figure 3: MCMC simulations of logistic regression weights for Dataset 0, and corresponding kernel density plot, assuming an improper uniform prior. Subset of 1,000 out of 10,000 total iterations shown.

4 Markov-Chain Monte Carlo

To evaluate the quality of the covariance estimates produced by our method, we used Markov-Chain Monte Carlo (MCMC) as a benchmark for the "true" distribution of the logistic regression weights (w_0, w_1, w_2). We used the `R` package `MCMCpack` with an improper uniform prior, 10,000 iterations, and burn-in rate of 1,000 iterations. We also compared to an MCMC simulation with a normal prior on the weights $\mathcal{N}(\mathbf{0}, 1000I)$ and found similar results. Figures 1 and 2 show the progression of the MCMC algorithm assuming each prior.

To visualize the joint distribution of the logistic regression weights, we plot the MCMC results for the values of w_1 and w_2. In addition, we fit a kernel density to the MCMC simulated weights as a proxy for the contour plot of this empirical distribution, using the `R` package `MASS`, which is shown in Figure 3.

In order to evaluate the accuracy of the covariance matrix of $\mathbf{w}$ given by MFVB logistic regression, which is 3-dimensional, we overlay its contour plot with the MCMC kernel density plot, for Dataset 0 along the dimensions of w_1 and w_2. The contour plot of MFVB logistic regression for the logistic regression weights is bivariate normal $\mathcal{N}(\tilde{\mathbf{w}}, \tilde{\mathbf{V}})$, where $\tilde{\mathbf{w}}, \tilde{\mathbf{V}}$ are the MFVB parameters excluding the components involving w_0. Figure 4 shows the contour plot of MFVB and the kernel density of MCMC on Dataset 0. Since MFVB is known to underestimate the covariance of the underlying distribution, we expect the kernel density of MCMC to be more spread out than the contour plot of MFVB logistic regression, which is what we observe.

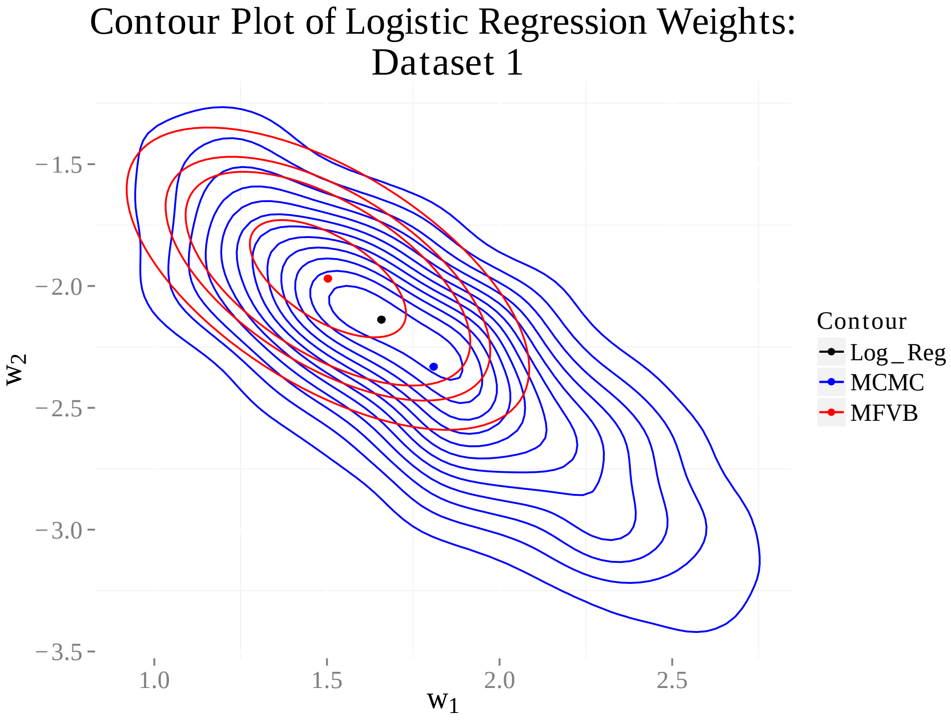

Figure 4: Comparison of logistic regression point estimate, kernel density plot of MCMC simulations, and posterior density of MFVB logistic regression (vectorized function), for Dataset 0. Contours of MFVB logistic regression indicate 50%, 90%, 95%, and 99% confidence intervals for the bivariate normal $\mathcal{N}(\mathbf{w}_N, \mathbf{V}_N)$. Mean values for MCMC and MFVB are also included.

Dataset	Logit Train	Logit Test	MCMC Train	MCMC Test	MFVB Train	MFVB Test
0	0.8700	**0.8580**	0.8700	**0.8580**	0.8600	0.8570
1	0.7500	0.8450	0.7500	**0.8460**	0.7400	0.8450
2	0.5500	**0.4890**	0.5100	**0.4890**	0.5400	0.4750
3	0.6900	**0.7420**	0.6900	**0.7420**	0.7200	0.7120
4	0.8600	**0.8050**	0.8600	**0.8050**	0.8600	0.8040
5	0.9400	0.9290	0.9400	**0.9300**	0.9400	0.9260
6	0.7400	0.7130	0.7400	**0.7140**	0.6900	0.6840
7	1.0000	0.9760	1.0000	***	1.0000	**0.9780**
8	0.9200	**0.8670**	0.9200	**0.8670**	0.9300	0.8640
9	0.5200	0.3980	0.5100	0.4020	0.4600	**0.4130**
10	0.6700	0.6690	0.6700	**0.6700**	0.6300	0.6560

Table 1: Training and test set accuracy for logistic regression, MCMC logistic regression, and MFVB logistic regression on all datasets. The highest out-of-sample accuracy score is highlighted for each dataset.
*** : The training dataset was completely separable; therefore the posterior covariance matrix was zero and the precision matrix was undefined. Thus, we did not obtain MCMC predictive probabilities for this dataset.

5 Computational Results

We compare the training and test set accuracy for all of the simulated datasets, which are summarized in Table 1. The predictive function for standard logistic regression is simple and the predictive function for MFVB logistic regression has already been discussed in Section 3. However, the predictive function for MCMC is new, and we will describe here our method. To predict the label $P(y=1|\mathbf{x})$ for MCMC, we use the predictive function for MFVB logistic regression, except substituting the empirical mean $\bar{\mathbf{x}}$ and empirical covariance $\hat{\Sigma}$ from the MCMC samples in place of the MFVB mean $\mathbf{w}_N$ and standard deviation $\mathbf{V}_N$.

Overall, we see that MCMC yields the highest out-of-sample accuracy in all cases except for Dataset 9. (In Dataset 7, no result for MCMC was reported because the data was separable.) Logistic Regression is the second strongest method, and for most datasets logistic regression yields out-of-sample accuracy results matching MCMC. Here, MFVB performs worse than the other methods, which is expected because MFVB requires the most restrictive set of assumptions: namely that the joint distribution factorizes into independent marginals. Because MFVB is an approximate method, it much runs faster than MCMC, but as a result may sacrifice some predictive power. For larger problems with many observations n and high-dimension p, this tradeoff between computational complexity and performance may become more significant.

Since the datasets are in 2-dimensions, we also present the decision boundaries for logistic regression, MCMC, and MFVB for selected datasets in Figures 5-10. We note that for many datasets, the decision boundaries are nearly identical for MCMC and standard logistic regression, while MFVB tends to be different from the other two.

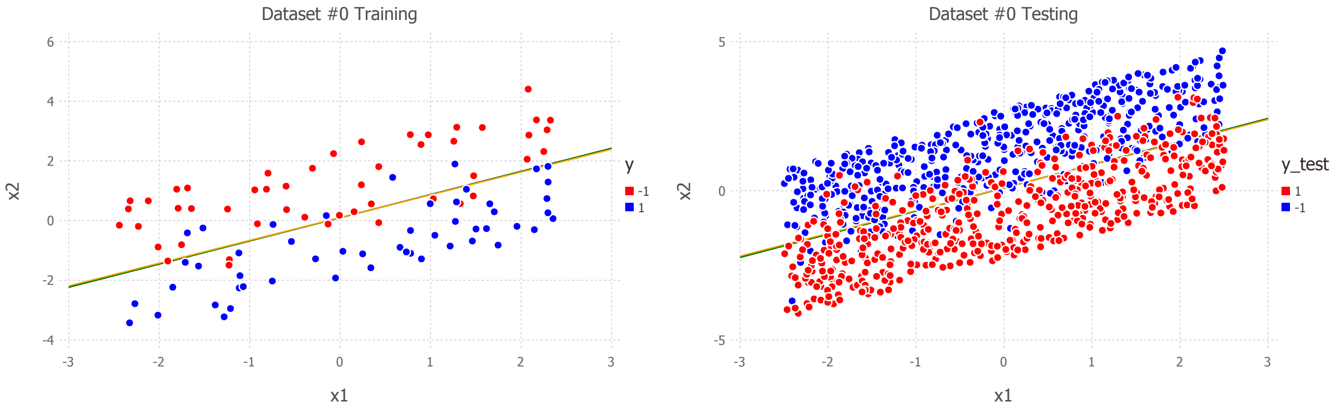

Figure 5: Dataset 0 decision boundary plots. **MCMC: Green**, **Log Reg: Black**, **MFVB: Orange**.

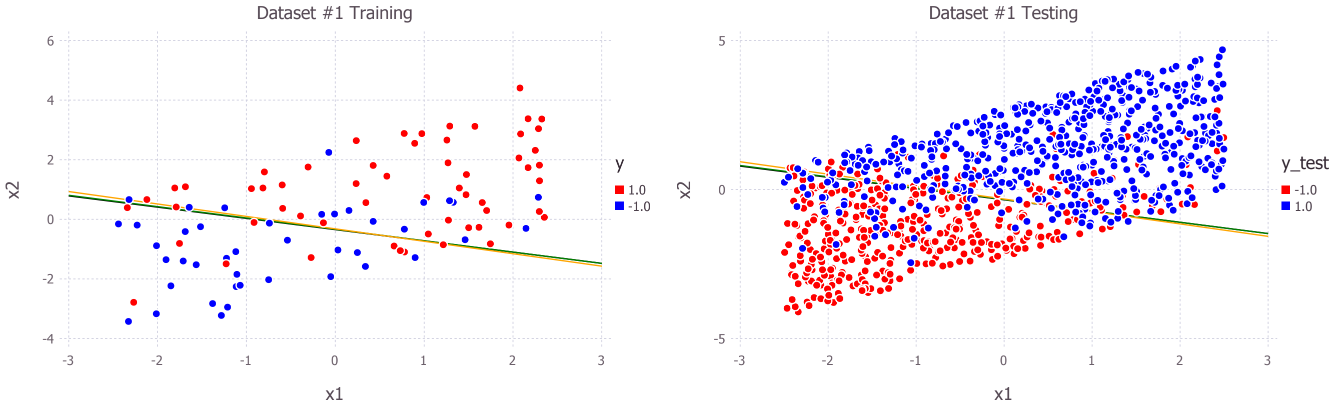

Figure 6: Dataset 1 decision boundary plots. **MCMC: Green**, **Log Reg: Black**, **MFVB: Orange**.

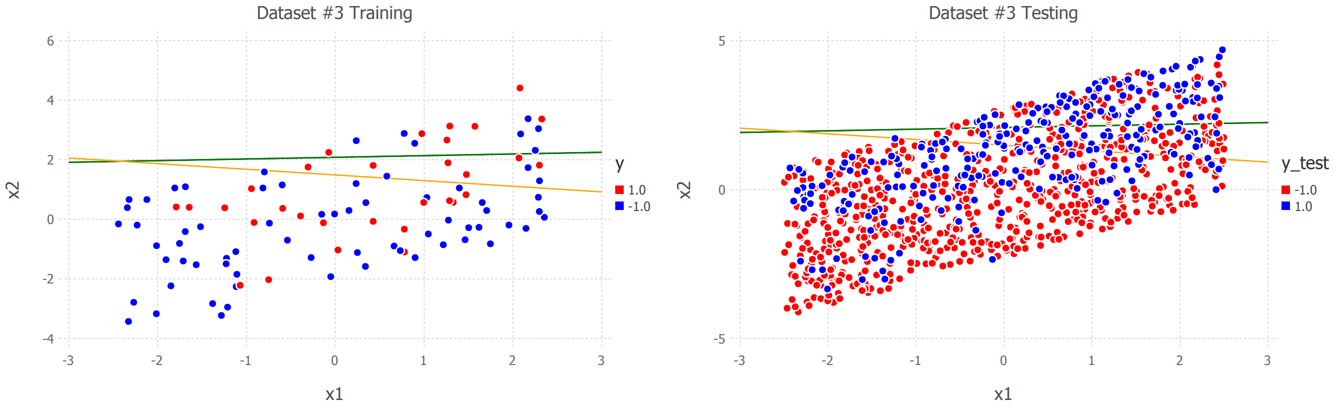

Figure 7: Dataset 3 decision boundary plots. **MCMC: Green**, **Log Reg: Black**, **MFVB: Orange**.

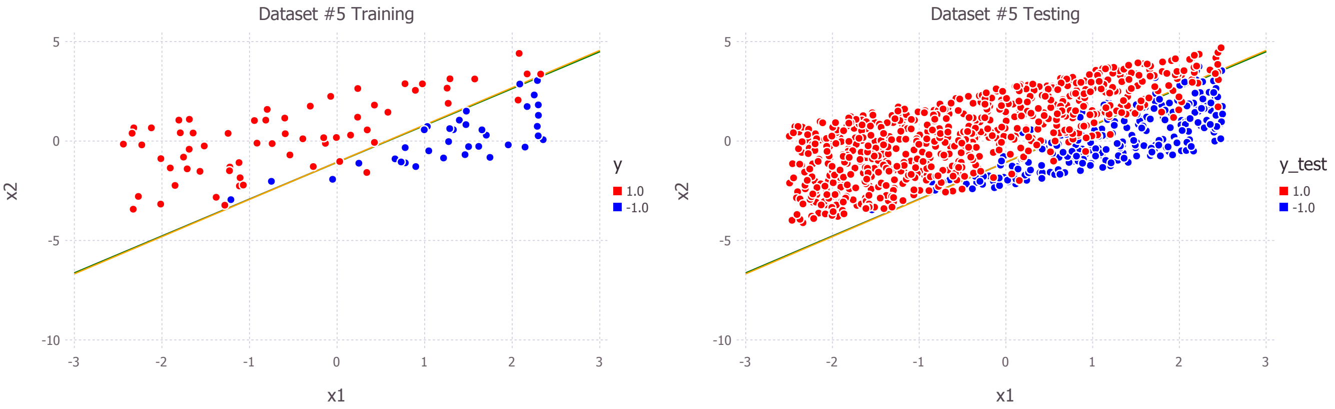

Figure 8: Dataset 5 decision boundary plots. **MCMC: Green**, **Log Reg: Black**, **MFVB: Orange**.

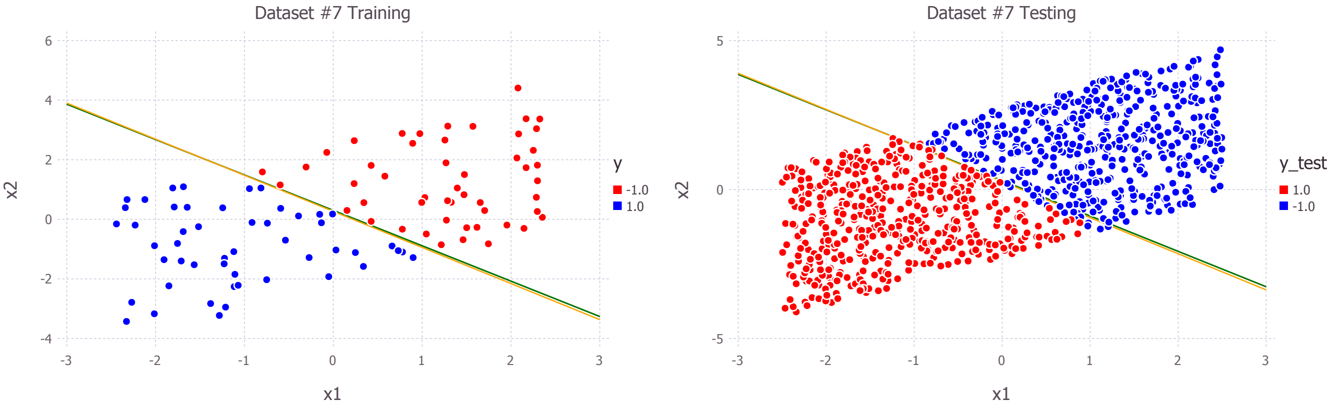

Figure 9: Dataset 7 decision boundary plots. **MCMC: Green**, **Log Reg: Black**, **MFVB: Orange**.

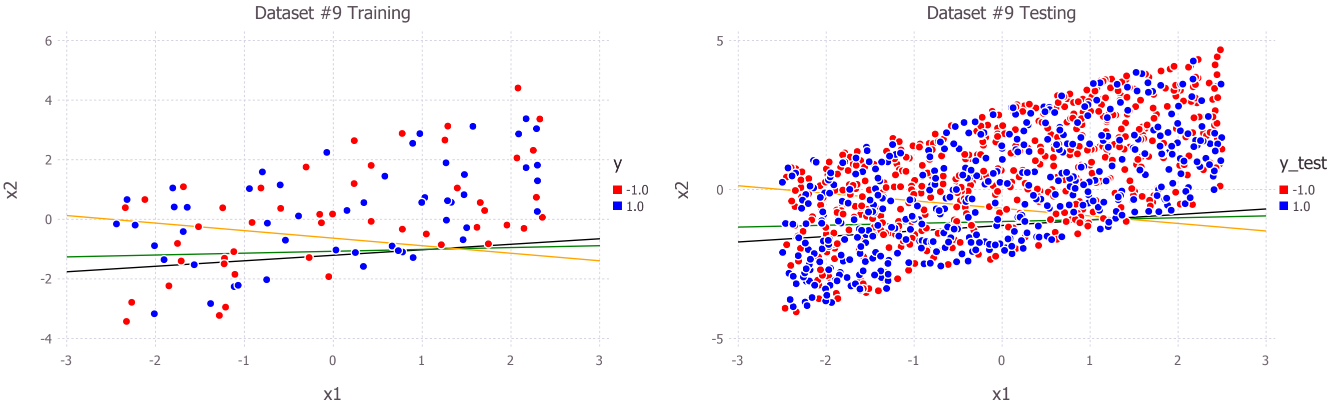

Figure 10: Dataset 9 decision boundary plots. **MCMC: Green**, **Log Reg: Black**, **MFVB: Orange**.

6 Future Work

A recent paper by Giordano, Broderick and Jordan (2015) proposes a linear correction term derived analytically to improve the covariance estimate for MFVB [2]. The new proposed method is called Linear Response Variational Bayes (LRVB). In the beginning of this project, we considered applying this new method on logistic regression and comparing its performance to the benchmark of MFVB and regular logistic regression. However, we ran out of time to complete this phase, because implementing MFVB for logistic regression was challenging and generating the plots for the MCMC vs MFVB comparison was difficult. In the future, we are planning to experiment with LRVB applied to variational Bayesian logistic regression, because we have majority of the most difficult work completed and now we have the source code for this method. From this research, we hope to develop an improved version of MFVB logistic regression.

So far, our results support that this is a promising idea. While MCMC and logistic regression yield comparable out-of-sample accuracy results, there is a noticable gap between MCMC and MFVB logistic regression. In addition, we observe from the contour plots that MFVB logistic regression underestimates the covariance of the $\mathbf{w}$, so we expect the magnitudes of the predicted probabilities to be different as a result. In addition, on small-scale datasets these methods run very fast, and they will be able to scale up to larger datasets and higher-dimensions while remaining computationally tractable.

References

[1] Jan Drugowitsch. Variational bayesian inference for linear and logistic regression. *arXiv preprint arXiv:1310.5438*, 2013.

[2] Ryan Giordano, Tamara Broderick, and Michael Jordan. Linear response methods for accurate covariance estimates from mean field variational bayes. *arXiv preprint arXiv:1506.04088*, 2015.

[3] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[4] Andrew D Martin, Kevin M Quinn, Jong Hee Park, et al. Mcmcpack: Markov chain monte carlo in r. *Journal of Statistical Software*, 42(9):1–21, 2011.

[5] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer Science & Business Media, 2009.

[6] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.